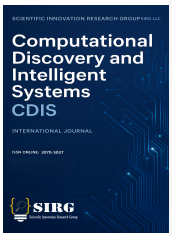

Computational Discovery and Intelligent Systems
CDIS
ISSN: 3070-5037/© 2026 CDIS. (CC BY 4.0) License.
Journal Homepage
https://pub.scientificirg.com/index.php/CDIS

EchoTwin Privacy Aware Personalized Email Reply Generation with Low Rank Adaptation

Thomas Samuel [a,1] **, Mohammed Abdelwahab** [a] **, Hanaa Atwa**[a] **, Anasimon Samir** [a]**, and Carol Faried** [a]

[a] *Faculty of Computer Science, Nahda University, Beni-Suef City, 62511, Egypt;*

E-mails: t.samuel0712@nub.edu.eg, mohamed.abdelwahab@nub.edu.eg, hanaa.atwasayed@gmail.com, a.samir2982@nub.edu.eg, and c.faried0147@nub.edu.eg

ABSTRACT

Email continues to serve as a primary medium for both professional and personal communication. Although contemporary assistance systems, including Smart Reply and Smart Compose, enhance writing efficiency, they generally do not explicitly model or adapt to an individual's unique communication patterns. This paper presents EchoTwin, an AI-assisted email-reply generation system that produces personalized draft replies aligned with a user's historical communication patterns. EchoTwin builds on a decoder-only Transformer large language model and applies Low-Rank Adaptation (LoRA) for parameter-efficient personalization: a lightweight adapter is trained for each user while a single base model is shared across the population, avoiding the cost of an independent model per user. The system integrates with Gmail through Google OAuth 2.0 and the Gmail API, keeping a human reviewer in the loop to approve, edit, or reject every draft before it is sent. The approach was evaluated on a curated subset of the Enron Email-Reply dataset, comprising 2,221 email-reply pairs from ten users, comparing personalized generation against the base model on 223 held-out samples using lexical similarity (ROUGE), semantic similarity, corpus-level style similarity, and stylometric analysis. The personalized adapters achieved statistically significant gains of 8.10% in ROUGE-1, 35.65% in ROUGE-L, and 5.64% in corpus-level style similarity, plus a 25.26% reduction in mean stylometric distance (Wilcoxon signed-rank test, $p < 0.05$). Sentence-BERT similarity to the reference reply decreased by 10.29%, reflecting the expected trade-off between reference-based similarity and individual style adaptation. These prototype-level results indicate that user-specific LoRA adapters can improve measured stylistic alignment in personalized email-reply generation while keeping the user in control of the final output.

PAPER INFORMATION

HISTORY

Received: 15 April 2026
Revised: 18 June 2026
Accepted: 30 July 2026
Online: 29 August 2026

MSC

68T07; 68R10; 94A60; 68M15

KEYWORDS

EchoTwin;
Privacy-Aware Architecture;
Personalized Email-Reply;
Low-Rank Adaptation;
Large Language Models.

1 INTRODUCTION

Email is one of the most common forms of digital communication in professional, academic, and personal settings, and it occupies a substantial share of the knowledge worker's day. Earlier productivity studies estimate that email-related activities account for roughly 28% of the workweek for interaction workers [1], and a typical full-time professional receives more than 120 messages per day [2], which turns email management into a considerable productivity challenge. Several AI-powered

[1]Corresponding author at Faculty of Computer Science, Nahda University, Beni-Suef City, 62511, Egypt;
E-mail: t.samuel0712@nub.edu.eg

writing systems have been built to reduce the effort involved in composing a reply. Smart Reply and Smart Compose, for instance, offer automated reply suggestions and real-time writing assistance that measurably improve communication efficiency [3, 4]. These systems are practical and already deployed at scale, but they are primarily built for general assistance and are not designed to adapt explicitly to the vocabulary, tone, or stylistic preferences of an individual user.

Generating personalized email responses that reflect an individual's communication style, while remaining computationally efficient and scalable, is still an open research problem. Learning user-specific patterns from a handful of past interactions is difficult, and training a separate large language model for every user is infeasible given the associated computational and storage costs. Parameter-efficient personalization methods are therefore needed to preserve individual writing style while retaining a shared base model. To address this challenge, this paper proposes a privacy-aware, AI-assisted email-reply generation system named EchoTwin, built on user-specific Low-Rank Adaptation (LoRA) adapters attached to a decoder-only Transformer large language model [5, 6]. EchoTwin combines a shared base language model with lightweight, user-specific LoRA adapters learned from historical email-reply pairs. This parameter-efficient design enables personalized email generation without requiring a dedicated model per user, improving scalability while preserving personalization capability.

Beyond personalization, the EchoTwin architecture is built around privacy-by-design principles, including secure authentication through Google OAuth 2.0, least-privilege access control, secure token management, data minimization, encrypted storage of sensitive artifacts, and user-controlled revocation of permissions [7, 8, 9, 10]. These design choices make practical deployment feasible while limiting the privacy risks that arise from processing personal email content.

The central aim of this work is to examine whether user-specific LoRA adapters can generate more personalized email replies than a non-personalized base language model. To this end, EchoTwin is evaluated with a suite of complementary automatic metrics spanning lexical similarity, semantic similarity, corpus-level style similarity, and stylometric analysis, applied to a curated subset of the Enron Email-Reply dataset.

The main contributions of this work are summarized as follows.

- A LoRA-adapter-based, user-specific personalization approach that enables parameter-efficient adaptation without altering the underlying shared language model.
- An empirical study of personalized email-reply generation on the Enron Email-Reply dataset, covering ten users and 2,221 email-reply pairs.
- A comprehensive evaluation methodology for personalized email generation that combines lexical, semantic, stylistic, and stylometric analyses.
- Experimental evidence indicating that user-specific LoRA adaptation achieves stronger stylistic alignment and writing-style consistency than a non-personalized baseline, supported by statistical significance testing.

The remainder of the paper is organized as follows. Section 2 reviews related work on AI-assisted email response generation, personalized language modeling, privacy-preserving personalization, and evaluation benchmarks. Section 3 describes the problem statement, key challenges, and personalization objective. Section 4 presents the architecture and design of the proposed EchoTwin system. Section 5 describes the dataset and data preparation pipeline. Section 6 details the machine learning methodology and reply-generation model. Section 7 describes the experimental setup and evaluation protocol. Section 8 reports the experimental results. Section 9 discusses limitations and directions for future work, and Section 10 concludes the paper.

2 RELATED WORK

Recent advances in large language models (LLMs) have substantially improved generated-text quality and enabled new forms of personalized content generation. Published work between 2022 and 2026 reflects a transition from generic language generation toward user-aware systems that incorporate personalization, feedback, retrieval mechanisms, and privacy considerations. Maintaining consistent user-specific behavior nevertheless remains difficult, particularly when user data is limited or privacy constraints restrict centralized data collection.

The use of AI to support email composition is now common in everyday communication tools. Smart Reply introduced neural response-suggestion techniques that generate short candidate replies for incoming emails, allowing users to respond efficiently with minimal interaction [3]. Smart Compose extended this idea by providing real-time writing suggestions and predictive text during composition [4]. These systems have delivered measurable productivity gains and have been deployed at scale, with the primary goal of increasing writing throughput, reducing drafting effort, and accelerating communication. However, because they are built to serve a broad user population, learning user-specific communication behavior is not their primary design objective, and their suggestions may not always align with an individual's stylistic preferences, vocabulary, or response habits. This observation does not diminish the effectiveness of existing email assistance systems; rather, it

points to a complementary research direction that personalizes language generation, that is, adapting generated content to an individual's communication behavior while preserving response quality and usability.

Personalization is widely recognized as a key direction for improving the relevance and usefulness of text generated by large language models. Existing approaches span parameter-efficient fine-tuning, retrieval-augmented generation, and user-specific adaptation techniques. FedPC enables preference-aware personalization in a federated language-generation setting and reports improved personalized text generation at low memory cost [11]. Retrieval-based personalization has also received considerable attention: optimization-based retrieval methods show that retrieving more relevant user documents improves personalized text-generation performance [12], and parameter-efficient personalization methods demonstrate that adapting language models with user-specific preferences can substantially improve personalized generation while remaining computationally efficient [13]. A recent comparative study contrasting retrieval-augmented generation (RAG) and parameter-efficient fine-tuning (PEFT) at a peer-reviewed venue shows that combining retrieval and adaptation can outperform either technique in isolation for privacy-preserving personalization [13]. Democratizing large language models through personalized PEFT has likewise been shown to improve individualized response quality while updating only a small fraction of model parameters per user [14], and collaborative data-refinement strategies such as Persona-DB improve personalized response prediction by structuring user history more effectively before adaptation [15]. A broad survey of personalized generation in the large-model era further organizes this space into representation, retrieval, and adaptation-based paradigms and highlights email and messaging assistance as an emerging application area [16].

Personalized language generation is closely related to text style transfer, low-resource adaptation, and parameter-efficient fine-tuning research. In existing style-transfer work, changing stylistic features while preserving semantic meaning remains challenging, especially with limited user-specific data. Personalization in few-shot settings faces a similar trade-off between adaptation and generalization, since models trained on small user datasets may fail to learn coherent communication patterns or may overfit to a small number of examples. Parameter-efficient fine-tuning methods such as LoRA offer a practical alternative to fine-tuning an entire model, reducing computational and storage costs and mitigating catastrophic forgetting, although they still involve a trade-off between adaptation capacity and personalization fidelity. These findings motivate the use of user-specific LoRA adapters in EchoTwin and underline the need to evaluate personalization under realistic, low-resource data conditions.

Because personalization often relies on sensitive user information, privacy-aware adaptation techniques for large language models have received growing attention. RAPT combines local differential privacy with prompt tuning to protect user information while retaining competitive task performance [17]. PrivateLoRA introduces an edge-cloud framework for privacy-aware LoRA adaptation that reduces communication overhead while preserving personalization capability [18]. PrivacyRestore targets inference-time privacy by removing sensitive information before generation and restoring it afterward, with promising results in privacy-sensitive domains [19]. More recent work extends privacy-preserving adaptation to the federated setting: FedEx-LoRA proposes exact aggregation for federated and efficient fine-tuning of large models, correcting the approximation error introduced by naive federated averaging of low-rank updates [20], while selective decoupled federated LoRA schemes address client heterogeneity under differential-privacy constraints by separating structurally mixed updates before aggregation [21]. A related line of work re-examines the interaction between LoRA's low-rank structure and differential-privacy noise, showing that gradient coupling between the two low-rank factors can amplify privacy noise and degrade utility unless the update directions are explicitly decoupled [22]. Together, these studies confirm that personalization and privacy protection are not mutually exclusive, although most existing approaches address individual privacy mechanisms in isolation rather than full user-facing communication systems.

Several benchmarks have been proposed in recent years to support the evaluation of personalized language-generation systems. LaMP is a standard benchmark for personalized language modeling across a range of generation and classification tasks [23]. LongLaMP extends this benchmark to long-form generation scenarios that demand stronger stylistic consistency and user adaptation [24]. PERSONAMEM addresses dynamic user profiling and assesses how well a model adapts to user-preference drift over time [25]. PREF is a reference-free evaluation framework for personalized text generation that correlates well with automatic measures of personalization quality [26]. PRISP studies few-shot personalization with privacy preservation and reports effective adaptation under limited-data conditions [27]. Together, these benchmarks highlight three recurring challenges for personalized language generation: effective personalization with limited user data, adaptation to evolving user behavior, and reliable automatic evaluation of personalization quality. These challenges directly motivate the design of EchoTwin and the multi-perspective evaluation methodology adopted in this work, which combines lexical similarity, semantic similarity, corpus-level style similarity, and stylometric consistency.

The reviewed literature reports substantial progress in personalized language generation through retrieval augmentation, parameter-efficient fine-tuning, reinforcement learning from human feedback, privacy-aware adaptation, and benchmark development. Despite these advances, three challenges recur across existing studies: effective personalization with limited user data, privacy preservation during personalization, and reliable evaluation of personalization quality beyond traditional lexical similarity metrics. Most existing work addresses only a single aspect of personalization, such as retrieval models, adaptation techniques, privacy frameworks, or benchmark design, with comparatively little attention paid to personalized email-reply generation systems that combine user-specific adaptation with human supervision inside a

real-world communication workflow.

This gap motivates the development of EchoTwin, a personalized email-reply generation framework built on user-specific LoRA adaptation with human-in-the-loop supervision. The central question explored in this work is whether parameter-efficient, user-specific personalization can improve stylistic alignment and reply quality while keeping the user in control of the generated response. To evaluate this question, EchoTwin employs a multi-perspective evaluation framework consisting of lexical similarity, semantic similarity, corpus-level style similarity, and stylometric analysis, summarized in **Table 1**.

Table 1: Summary of Related Work

Research Area	Representative Studies (Dataset/Benchmark)	Key Contributions	Relation to EchoTwin
AI-Assisted Email Writing	Smart Reply [3] (Gmail logs); Smart Compose [4] (Gmail logs)	Improve writing efficiency via automated reply suggestions and predictive text; deployed at scale	EchoTwin targets personalized email-reply generation rather than general-purpose writing assistance
Personalized Language Generation	FedPC [11] (federated dialogue corpora); Personalized-RLHF [28]; Retrieval-Based Personalization [12] (LaMP); RAG–PEFT Comparison [13] (LaMP); Democratizing PEFT [14] (LaMP); Persona-DB [15] (dialogue history); PGen Survey [16] (multi-domain)	Show that user-specific adaptation, preference learning, retrieval augmentation, and parameter-efficient fine-tuning improve personalization quality	EchoTwin uses user-specific LoRA adapters to personalize email replies while retaining computational efficiency
Privacy-Preserving Personalization	RAPT [17]; PrivateLoRA [18]; PrivacyRestore [19]; FedEx-LoRA [20]; SDFLoRA [21]; Federated LoRA–DP Analysis [22] (all: synthetic/federated NLP corpora)	Protect user information during training or inference while supporting personalization, including federated LoRA aggregation	EchoTwin combines personalization with privacy-conscious deployment practices in an email communication workflow
Personalized Generation Evaluation	LaMP [23]; LongLaMP [24]; PERSONAMEM [25]; PREF [26]; PRISP [27] (all: purpose-built personalization benchmarks)	Provide benchmarks and evaluation frameworks for personalized generation, dynamic user adaptation, and privacy-aware personalization	EchoTwin applies lexical, semantic, stylistic, and stylometric evaluation to assess personalization quality
Research Gap	Studies above, individually	Significant progress on isolated aspects of personalization, privacy, or evaluation	EchoTwin addresses this gap through user-specific LoRA personalization, human-in-the-loop supervision, and comprehensive evaluation for personalized email-reply generation, on the Enron Email-Reply dataset [29, 30]

3 PROBLEM STATEMENT AND KEY CHALLENGES

This work studies whether user-specific personalization can increase the quality of AI-generated email replies while remaining under human oversight. Modern language models produce fluent responses, but they tend to generate generic output that does not match an individual's communication preferences. This limitation is especially relevant for email, where tone, word choice, and reply structure vary widely across users.

Personalizing email generation involves several challenges.

- **Generic responses and limited user adaptation.** General-purpose language models are trained to produce broadly

relevant responses. The resulting drafts can be logically coherent yet inconsistent with an individual's communication patterns; personalizing at the user level is difficult, particularly when only limited user data is available.

- **Limited personalized training data.** Personalization systems typically operate in low-data regimes, since many users have only a small number of past email replies available for adaptation. The system must therefore learn user-specific communication patterns from relatively short datasets without overfitting or memorizing training examples.
- **Human oversight and reliability.** AI-generated email drafts can contain incorrect assumptions, missing facts, or stylistic flaws. Generated responses should support the user rather than replace human judgment, and every generated draft remains subject to human review before it is sent.
- **Evaluation of personalized generation.** Evaluating personalization is inherently difficult because many valid answers can exist for the same email. Evaluation should therefore combine semantic relevance, similarity to a user's communication patterns, and, where feasible, human judgment of answer quality.

A side-by-side comparison of the traditional manual email workflow and the EchoTwin-assisted workflow is presented in **Figure 1**.

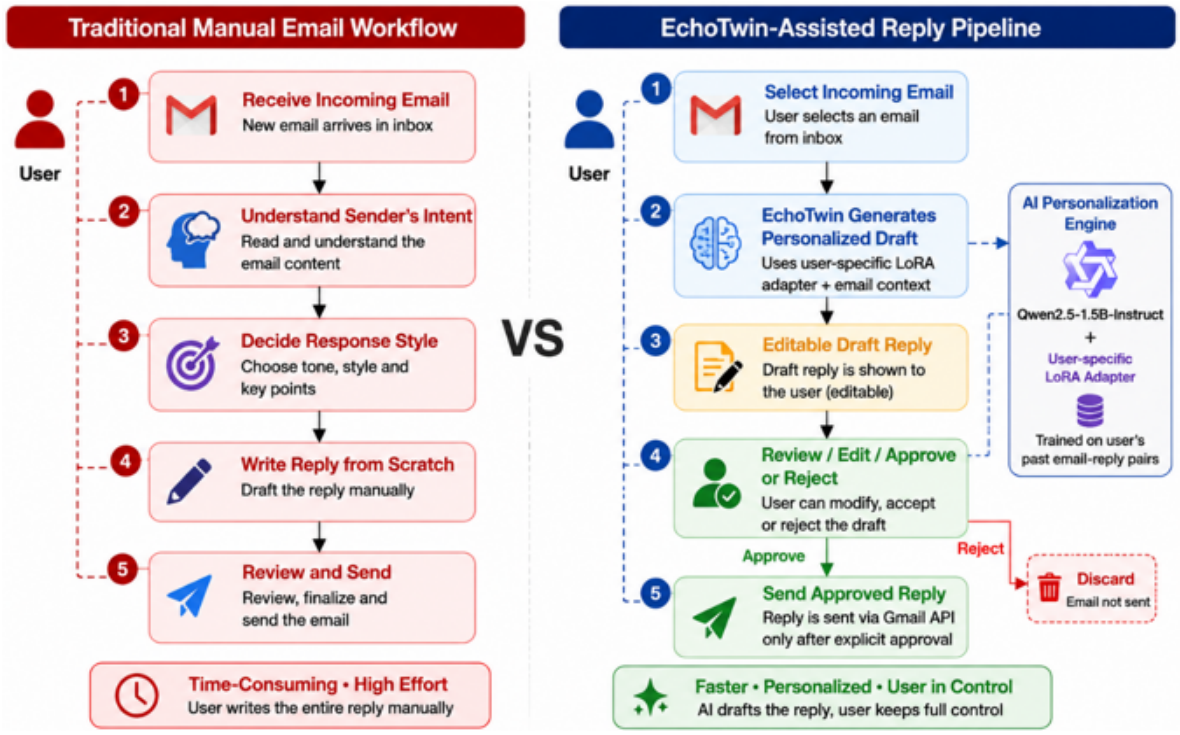

Figure 1: Traditional email workflow versus the EchoTwin-assisted reply pipeline

In a typical email workflow, the user receives an incoming message, interprets the sender's intent, decides on an appropriate response style, and writes the answer from scratch, a process that becomes time-consuming when email volume is high. EchoTwin instead provides the user with a personalized draft reply based on the selected email and a user-specific LoRA adapter trained on past email-reply pairs. The draft is presented in an editable format so that the user can review, modify, accept, or reject the response before it is sent, an approach intended to reduce drafting effort while preserving user control and accountability through explicit approval prior to transmission.

3.1 Personalized Reply Generation Objective

EchoTwin takes an incoming email and a target user identifier and produces a personalized reply draft that is relevant to the content of the email and consistent with the user's writing style. Unlike general-purpose language models, the proposed approach customizes reply generation by training user-specific Low-Rank Adaptation (LoRA) adapters on historical email-reply pairs.

Rather than optimizing an explicit, hand-crafted objective, personalization is learned directly through parameter-efficient fine-tuning. At inference time, the shared foundation model is combined with the matching user-specific LoRA adapter to generate responses that reflect the user's communication style while preserving the semantic intent of the incoming email. The effectiveness of this personalization approach is evaluated empirically through lexical similarity, semantic similarity, corpus-level style similarity, stylometric analysis, and statistical significance testing.

3.2 Main Research Challenges

Developing personalized email-reply generation systems raises challenges that extend beyond traditional text generation. Unlike general-purpose language models, personalized systems must adapt generated responses to the communication behavior of individual users while preserving contextual relevance and response quality. This work addresses three major research challenges.

Challenge 1: User-specific personalization. Communication style, preferred vocabulary, response structure, and writing conventions vary substantially across users. Responses generated by generic language models are generally appropriate but tend to be inconsistent with individual communication styles. The core challenge is therefore how to adapt a shared language model to generate responses that are more coherent with a given user's past communication behavior while preserving contextual relevance.

Challenge 2: Parameter-efficient personalization. Personalizing large language models for individual users through full-model fine-tuning is computationally expensive, and maintaining a separately fine-tuned model per user is impractical given storage requirements, training cost, and deployment complexity. The associated research challenge is how to adapt effectively to each user by updating only a small subset of model parameters, which motivates the use of Low-Rank Adaptation for lightweight personalization without retraining the underlying language model in full.

Challenge 3: Evaluation of personalized email generation. Personalized text generation is difficult to evaluate because many valid responses can exist for a given email. Traditional lexical similarity measures alone cannot capture personalization quality: a generated reply may differ substantially from the reference response yet still exhibit communication characteristics consistent with the target user. A central challenge is therefore to design an evaluation methodology capable of assessing lexical similarity, semantic similarity, stylistic alignment, stylometric consistency, and, where possible, human judgment.

4 PROPOSED ECHOTWIN ARCHITECTURE

EchoTwin is a privacy-preserving, personalized email-reply generation framework built on a shared Transformer large language model combined with user-specific Low-Rank Adaptation (LoRA) adapters. Rather than fine-tuning a foundation model separately for every user, EchoTwin uses a single shared language model together with a lightweight, user-specific adapter, enabling parameter-efficient personalization while reducing computational cost, storage requirements, and deployment complexity relative to full-model fine-tuning [18, 13, 6]. **Figure 2** shows the proposed personalization pipeline. The workflow begins with the selection of an incoming email by the user. The email context is encoded and passed to the shared decoder-only Transformer large language model together with the corresponding user-specific LoRA adapter. The adapter introduces user-dependent parameters into the frozen base model for personalized response generation without altering the original model weights. The resulting draft is then presented to the user for review prior to explicit approval and transmission.

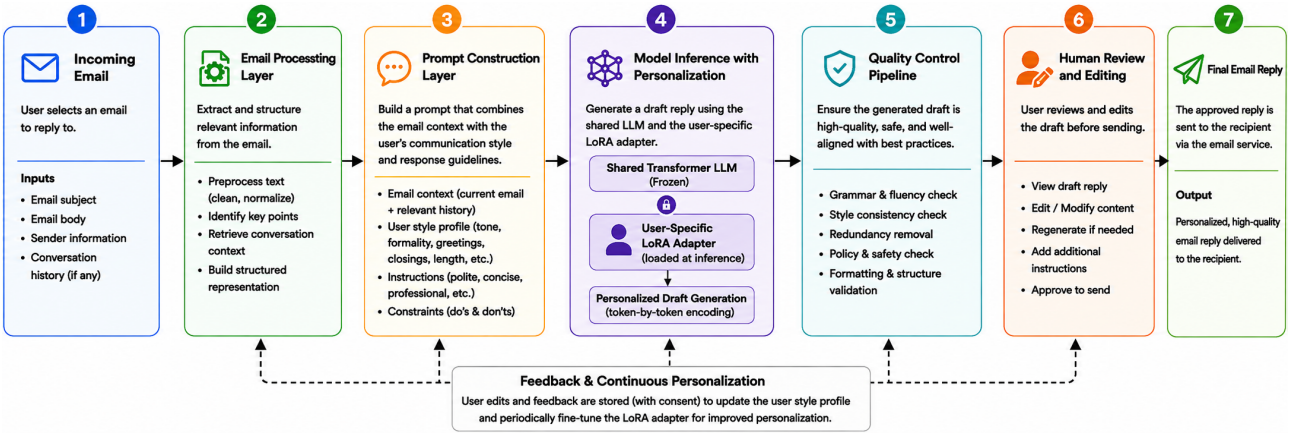

Figure 2: Proposed EchoTwin personalized email-reply generation framework

4.1 Email Processing Layer

The workflow begins when a user selects an email to reply to. The incoming message is parsed to extract the information needed for personalized generation, including the subject, message body, and any available conversation history. Rather than processing an entire mailbox, EchoTwin fetches only the selected email, following the principle of data minimization. The

extracted information is converted into a structured representation that serves as contextual input for the reply-generation process.

4.2 Context Representation and Response Generation

The email is represented in a structured form and passed to the personalized language model, which generates a draft response. Generation combines four sources of information: contextual understanding of the incoming email; conversational history, when available; linguistic knowledge encoded in the shared Transformer language model; and stylistic knowledge encoded in the selected LoRA adapter. EchoTwin is not a generic email assistant that produces uniform output for every user; generation is biased toward the vocabulary selection, sentence structure, greeting style, closing phrases, and overall communication patterns observed in each user's historical emails, with the goal of producing output that is semantically relevant and consistent with that user's writing style.

4.3 AI Personalization Layer

Per-user personalization is the central contribution of EchoTwin. The shared Transformer large language model provides the backbone for general-purpose language generation, and each user is associated with an independent LoRA adapter trained only on that user's historical email-reply pairs [5, 6]. At inference time, the appropriate adapter is dynamically loaded based on the authenticated user and merged with the shared model, so that only a small subset of trainable parameters is updated while the foundation model remains intact. This parameter-efficient approach offers several advantages: user-specific writing-style adaptation, reduced computational and storage requirements, scalable deployment across multiple users, and preservation of a common knowledge base shared by all users. Because each adapter is trained independently, personalization knowledge remains confined to a single user, limiting unintended cross-user influence during adaptation.

4.4 Human-in-the-Loop Validation

EchoTwin follows a human-in-the-loop workflow that treats artificial intelligence as an assistive technology rather than an autonomous decision maker. Instead of sending replies automatically, the system presents a personalized draft for review. Users may edit the generated reply, request an alternative response, accept the draft, or reject it entirely; the final message is sent only after explicit user approval. This design is intended to reduce the likelihood of inaccurate, misleading, or contextually irrelevant output reaching a recipient, and to preserve user accountability and control over what is sent. It is an architectural safeguard rather than a validated outcome: the experiments reported in this paper evaluate the personalized language generation component alone, offline and with automatic metrics, and do not measure how the review step performs in practice, including user satisfaction, editing effort, factual accuracy of drafts, or safety of the content a user might approve. These questions require a separate, user-facing study.

4.5 Privacy and Security Design

Privacy preservation is embedded into the architecture rather than added as a separate deployment feature. EchoTwin follows the principles of data minimization, least-privilege access, and user-controlled authorization consistent with current security recommendations [7, 8, 9, 10]. The framework incorporates several privacy mechanisms: OAuth 2.0 authorization with Proof Key for Code Exchange (PKCE) for secure authentication [7, 31]; secure backend token management that avoids exposing access tokens to client applications; encrypted communication over HTTPS/TLS [9]; isolated user-specific LoRA adapters that prevent knowledge leakage across users; retrieval of email content only when required for personalized reply generation; and revocable user authorization that allows access to be terminated at any time. Lightweight adapter parameters are stored individually per user, avoiding cross-user interference while keeping the foundation model shared, which enables scalable personalization without exposing one user's communication patterns to another user's adapter.

5 DATASET AND DATA PREPARATION

This section describes the dataset source, user-selection process, preprocessing pipeline, dataset partitioning, and LoRA-based personalization procedure used in EchoTwin. The objective of this phase was to transform raw email-reply records into a curated dataset suitable for user-specific personalization and experimental evaluation.

5.1 Dataset Source

EchoTwin was trained and evaluated using the Enron Email-Reply dataset distributed on Kaggle [29], a curated subset of the original Enron email corpus [30]. The dataset contains 15,377 email-reply pairs extracted from real workplace correspondence and has been widely used in natural language processing research on email [29, 30]. Each record includes an incoming email and a human-written response, together with metadata such as subject information, sender and user identifiers, timestamps, and conversation context. The dataset is well suited to personalized reply generation because it associates email responses with individual users, enabling both user-level adaptation and evaluation.

Provenance and ethical use. The underlying Enron corpus originates from internal company email entered into the public record by the Federal Energy Regulatory Commission (FERC) during its 2001–2003 investigation of the Enron Corporation. Carnegie Mellon University subsequently curated and redistributed the corpus for research purposes [30]; that curation involved limited removal of some sensitive material identified during processing, but it does not constitute formal, systematic de-identification, and residual personal or sensitive information may remain in individual message bodies. Consistent with the norms under which the corpus and its derivatives are distributed, this study treats it as licensed for non-commercial academic research only, and the data were used exclusively for offline model training and evaluation rather than for any production, commercial, or user-facing deployment. The user identifiers used throughout this paper (e.g., `daren_farmer`, `kay_mann`) are the pseudonymous mailbox labels already present in the public dataset; they were not created, linked, or independently verified by the authors, and no attempt was made to re-identify, contact, or infer additional information about the individuals named in the corpus.

5.2 User Selection and Dataset Construction

The experiments used the Enron Email-Reply dataset from Kaggle [29], extracted from the original Enron email corpus [30]. A preprocessing pipeline was applied to improve data quality and prepare the dataset for personalized email-reply generation. The pipeline removed HTML tags, URLs, email addresses, forwarded-message blocks, duplicate records, auto-generated replies, and messages with insufficient reply content. Remaining records were reorganized into a structured format containing the incoming message, the corresponding reply, the user identifier, and contextual information used for training and evaluation. Users were retained only if they had enough valid email-reply pairs to support reliable user-level personalization and evaluation, and the selected users exhibited variation in email volume, communication context, and reply characteristics, which allowed personalized reply generation to be assessed across a range of writing styles. The final dataset contains 2,221 email-reply pairs from 10 selected users, with sample counts varying by user depending on the amount of valid data remaining after preprocessing. **Table 2** lists the selected users and the number of email-reply pairs available for the study, and **Figure 3** summarizes the complete pipeline used for dataset preparation, user selection, training, and evaluation.

Table 2: Selected Users and Available Email-Reply Pairs

User	Total Valid Email-Reply Pairs
daren_farmer	170
gerald_nemec	200
jeff_dasovich	250
kay_mann	250
louise_kitchen	220
matthew_lenhart	131
michelle_nelson	250
mike_maggi	250
sara_shackleton	250
tana_jones	250

The user-selection process was guided by the requirement for sufficient user-specific personalization data rather than by anticipated model performance. Training a dedicated LoRA adapter requires enough historical email-reply pairs per user to support training, validation, and test partitions, so users without enough valid pairs after preprocessing were excluded because they could not support reliable user-level personalization experiments. Selection was therefore based on data sufficiency rather than predicted performance or communication characteristics, and it took place prior to model training and evaluation, so it did not involve performance-based user selection.

Figure 3: EchoTwin dataset preparation, user selection, training, and evaluation pipeline

5.3 Dataset Distribution Analysis

The final dataset was divided into training, validation, and testing sets at the email-reply pair level within each user. Sample counts differ across users because the split is conditioned on the amount of valid data remaining after preprocessing. **Table 3** summarizes the per-user dataset distribution used in the experiments.

Table 3: Dataset Distribution by User

User	Training	Validation	Testing	Total
daren_farmer	136	17	17	170
gerald_nemec	160	20	20	200
jeff_dasovich	200	25	25	250
kay_mann	200	25	25	250
louise_kitchen	176	22	22	220
matthew_lenhart	104	13	14	131
michelle_nelson	200	25	25	250
mike_maggi	200	25	25	250
sara_shackleton	200	25	25	250
tana_jones	200	25	25	250
Total	**1,776**	**222**	**223**	**2,221**

Overall, the complete dataset contains 1,776 training samples, 222 validation samples, and 223 testing samples, for a total of 2,221 email-reply pairs. The validation set was used to monitor training behavior and select model checkpoints, while the test set was kept separate from training and used exclusively for final evaluation. The distribution of training, validation, and testing samples across the selected users is shown in **Figure 4**.

5.4 Dataset Fields

The dataset was preprocessed and reorganized into the fields listed in **Table 4** for use in the EchoTwin training and evaluation pipeline.

The model received the incoming subject, email body, and available context as input, and the ground-truth reply was used as the target response for both personalization and evaluation.

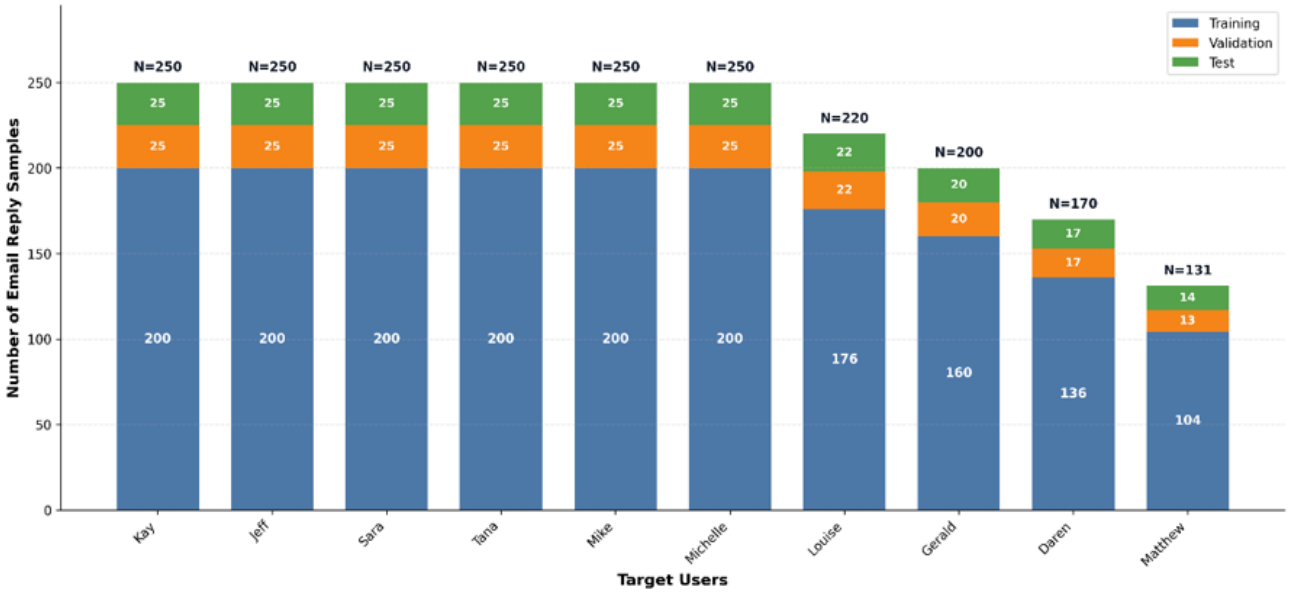

Figure 4: Training, validation, and testing sample distribution across selected users

Table 4: Main Fields in the Prepared Dataset

Field Name	Description
case_id	Unique identifier for each email-reply pair
sender_id	Email sender identifier
user_id	Target user associated with the reply
incoming_subject	Subject of the incoming email
incoming_body	Main content of the incoming email
reply_context	Additional conversation context, when available
ground_truth_reply	Original human-written reply
incoming_words	Number of words in the incoming email
reply_words	Number of words in the reply

5.5 Preprocessing Pipeline

Several preprocessing steps were applied to improve data quality and reduce noise before personalization. The cleaning pipeline removed incomplete or missing records, duplicate email replies, machine-generated answers, HTML tags and formatting artifacts, URLs and email addresses, extended signatures and forwarded-message blocks, overly short responses, and unsuitably long conversation threads.

5.6 Training and Validation Monitoring

Each user-specific adapter was trained for multiple epochs, with validation loss monitored throughout to avoid relying solely on the final training epoch. Training was carried out for four to five epochs, depending on the size of the user-specific dataset and the training configuration; validation loss was evaluated after every epoch, and the checkpoint with the best validation performance was retained for final evaluation. **Table 5** summarizes the training behavior observed across users.

For most users, the best validation performance was reached at epoch 2 or 3, suggesting that additional training epochs did not necessarily improve generalization.

5.7 Final Dataset Usage in EchoTwin

The prepared dataset supported both personalization and evaluation. User-specific LoRA adapters were trained on training samples, checkpoints were selected using validation samples, and the final result was evaluated on testing samples. For each test email, EchoTwin generated a personalized response using the corresponding user adapter and email context, and

Table 5: Training Epochs and Validation Monitoring

User	Training Epochs	Best Validation Epoch	Validation Tracking
daren_farmer	5	3	Yes
gerald_nemec	4	2	Yes
jeff_dasovich	4	2	Yes
kay_mann	4	2	Yes
louise_kitchen	4	2	Yes
matthew_lenhart	5	3	Yes
michelle_nelson	5	3	Yes
mike_maggi	5	3	Yes
sara_shackleton	5	3	Yes
tana_jones	5	3	Yes

the generated responses were compared with human responses using lexical similarity, semantic similarity, corpus-level style similarity, and stylometric analysis. The experiments were conducted in a controlled offline setup on a subset of Enron users, so the reported results should be considered prototype-level evidence of the effectiveness of personalization rather than proof of generalization to the entire Enron email population.

6 MACHINE LEARNING METHODOLOGY

This section describes the machine learning methodology adopted by EchoTwin for personalized email-reply generation. EchoTwin applies a parameter-efficient personalization approach using user-specific Low-Rank Adaptation (LoRA) adapters trained on historical email-reply pairs, with the goal of producing personalized reply drafts that are relevant to the incoming email and consistent with patterns observed in a user's previous responses.

6.1 Personalized Reply Generation Model

The core response-generation component of EchoTwin is built on a decoder-only Transformer large language model [5]. Personalization is performed using Low-Rank Adaptation (LoRA), a parameter-efficient fine-tuning method that adapts a model without updating its full parameter set [6]. Each selected user is trained with a dedicated LoRA adapter using that user's historical email-reply pairs, while a shared base model is retained across all users. During training, the original language model parameters remain frozen and only the adapter parameters are updated, which substantially reduces computational and storage requirements relative to full-model fine-tuning and allows personalized adaptation with limited user-specific data. At inference time, the shared base model and the corresponding user-specific adapter are loaded together to generate personalized email responses, enabling the system to adapt generated responses to historical user communication behavior while retaining the efficiency benefits of parameter-efficient fine-tuning.

6.2 User-Specific Conditioning

EchoTwin customizes reply generation using user-specific LoRA adapters rather than hand-crafted style classifiers or rule-based personalization logic. The model takes four inputs: email subject, email body, conversation context, and target user identifier. The corresponding personalized adapter is loaded according to the target user to generate the response. The personalization process does not rely on manually engineered stylistic features or explicitly computed user-profile scores; instead, user-specific communication patterns are learned implicitly through LoRA fine-tuning on historical email-reply pairs. This design supports efficient personalization through independent, user-specific adapters while retaining a shared base language model, with each adapter trained independently to model the communication patterns present in that user's historical data.

6.3 Training Strategy

Training used the standard causal language modeling objective applied to fine-tuning large language models, in which the model is optimized to predict target reply tokens from previously generated tokens and the input email content. A separate LoRA adapter was trained on each selected user's private training subset, and validation loss was monitored to select the best checkpoint for analysis. To support reproducibility, all experiments used a fixed random seed, identical training

hyperparameters, and a consistent train-validation-test split protocol across users. **Table 6** summarizes the training and reproducibility configuration used for all user-specific LoRA adapters.

Table 6: LoRA Training and Reproducibility Configuration

Parameter	Value
Base model	Decoder-only Transformer large language model
Personalization method	User-specific LoRA adapters
Adapter setup	One LoRA adapter per user
LoRA rank	16
LoRA alpha	32
LoRA dropout	0.05
Target modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
Number of training epochs	5
Learning rate	2e-4
Optimizer	AdamW
Per-device training batch size	1
Per-device evaluation batch size	1
Gradient accumulation steps	8
Effective training batch size	8
Weight decay	0.01
Warmup strategy	Warmup ratio = 0.03
Learning rate scheduler	Cosine
Evaluation strategy	Epoch-based evaluation
Save strategy	Epoch-based checkpoint saving
Logging strategy	Epoch-based logging
Checkpoint limit	2
Best checkpoint selection	Lowest validation loss
Metric for best model	eval_loss
Higher metric is better	False
Mixed precision	FP16 enabled
Data collator	DataCollatorForLanguageModeling
Masked language modeling	False
Hardware environment	Google Colab
GPU type	NVIDIA T4

For consistency and fair comparison, identical hyperparameter settings were applied to all users. Training was performed for five epochs, with validation loss monitored throughout, and the checkpoint with the lowest validation loss was selected for final evaluation. This strategy reduces the risk of overfitting and ensures that model selection is based on validation performance rather than on the final training epoch.

6.4 Reply Generation Pipeline

At inference time, the incoming email is converted into a structured prompt containing the subject, body, and any available conversation context. The corresponding user-specific LoRA adapter is loaded and merged with the shared decoder-only Transformer large language model, and the response is generated based on the email content and the patterns learned from the target user's historical email-reply pairs. The resulting output is post-processed to normalize formatting, remove generation artifacts, and produce a clean email structure before being used for evaluation and presented to the user for interaction.

6.5 User-Specific LoRA Personalization

EchoTwin adopts a user-specific personalization strategy in which each user has a dedicated LoRA adapter. This design offers several advantages: efficient personalization without full language model fine-tuning; reduced storage requirements relative to storing separate full models; independent adaptation across users, with no parameter sharing; and scalable deployment on top of a single shared base model. Because adapters are trained independently, personalization is isolated at the user level, which simplifies deployment and supports future extension to larger user populations.

6.6 Inference Configuration

During evaluation, the trained user-specific LoRA adapters were loaded together with the decoder-only Transformer large language model to generate personalized replies for unseen test samples [5, 6]. For each test email, the system received the email subject, body, available conversation context, and target user identifier. The corresponding LoRA adapter was selected and used to generate a personalized draft reply. Ground-truth replies were excluded from the input prompt and used only during evaluation, which prevents data leakage and ensures that generated responses depend solely on the incoming email content and the personalization patterns learned during training. This use of held-out test samples follows standard evaluation practice in machine learning and personalized language generation [23, 24]. **Table 7** summarizes the inference configuration used during reply generation.

Table 7: Inference Configuration Used in EchoTwin

Parameter	Value
Base model	Decoder-only Transformer large language model
Personalization	User-specific LoRA adapters
Input fields	Subject, incoming email, optional context
Output field	Generated reply
Adapter selection	Based on target user
Ground-truth reply in prompt	No
Maximum new tokens	180
Decoding strategy	Sampling
Temperature	0.75
Top-p	0.90
Top-k	Not used
Repetition penalty	1.12
Beam search	Not used
Stopping criterion	End-of-sequence token or maximum token limit

This configuration was chosen to balance relevance, fluency, and diversity while limiting repetitive or overly generic responses. Generated replies were subsequently evaluated using the automatic metrics described below.

6.7 Evaluation Methodology

Generated responses were evaluated using a set of complementary automatic metrics designed to assess both response quality and personalization performance. Multiple evaluation perspectives were used, rather than a single metric, because personalized email generation must preserve both contextual relevance and user-specific writing characteristics. The evaluation methodology comprises four complementary components.

- **Lexical similarity (ROUGE):** word-level overlap between generated and reference replies.
- **Semantic similarity (Sentence-BERT):** semantic similarity between generated and human-written replies, based on embedding cosine similarity [32].
- **Corpus-level style similarity:** the degree to which generated replies match the patterns observed in the target user's historical replies.
- **Stylometric analysis:** measurable characteristics of writing, including reply length, sentence structure, punctuation, and greeting and closing expressions, used to assess stylistic similarity to historical user replies, following established

authorship-attribution methodology [33, 34].

Combining lexical, semantic, and stylistic evaluation provides a multi-perspective assessment of personalized email-reply generation and enables the influence of user-specific personalization to be analyzed from complementary angles. All four components are automatic, reference-based or corpus-based metrics computed offline on held-out data; none of them involves human raters, and the evaluation accordingly does not measure user satisfaction, drafting or editing effort, factual correctness of generated content, or safety, each of which would require a separate study involving human participants.

7 EXPERIMENTAL SETUP

7.1 Implementation Environment

The EchoTwin prototype was implemented as a modular web-based system. The front end was built using React [35], and the back end was implemented using ASP.NET Core. Python was used to implement AI-related services with FastAPI [36], and data storage and management relied on Microsoft SQL Server [37]. The machine learning components were implemented using PyTorch and the Hugging Face Transformers framework [38]. The system uses Qwen2.5-1.5B-Instruct, a decoder-only Transformer large language model, as the shared base model [5]. User-specific personalization was performed with Low-Rank Adaptation (LoRA), in which lightweight adapter parameters were fine-tuned while the shared base model parameters remained frozen [5, 6]. Service orchestration and deployment were managed using Docker Compose. Gmail integration was secured with Google OAuth 2.0, PKCE, and the Gmail API [7, 8, 31]. OAuth credentials and access tokens were processed entirely on the backend and were never exposed to the frontend, consistent with secure credential management practices and privacy-by-design principles [9].

7.2 Experimental Dataset

The experiments used the preprocessed Enron Email-Reply dataset described in Section 5 [29, 30]. The final dataset contained 2,221 email-reply pairs from 10 selected users, divided into training, validation, and testing subsets. LoRA adapter fine-tuning was performed on the training data, validation data supported checkpoint selection and training monitoring, and testing data was reserved for final evaluation only. Further details on preprocessing, user selection, and dataset distribution are provided in Section 5.

7.3 Evaluation Procedure

The experiments assessed the effectiveness of user-specific LoRA personalization for email-reply generation. For each test email, two responses were generated: one produced by the decoder-only Transformer large language model without adaptation, and one produced with the corresponding user-specific LoRA adapter. Both responses were compared against the ground-truth human-written reply in the dataset. Results are reported on the complete held-out test set of 223 unseen email-reply pairs. This experimental design enables a direct comparison between non-personalized generation and user-specific personalization under identical evaluation conditions.

7.4 Evaluation Framework

The evaluation metrics and methodology used in this study are described in Section 6.7. The experimental evaluation compares personalized LoRA-based generation with the base model using lexical, semantic, stylistic, and stylometric measures, with the goal of quantifying both reply quality and user-specific personalization performance on the held-out test set.

7.5 Baseline Selection

To evaluate the effectiveness of user-specific personalization, EchoTwin was compared with the original base model (Qwen2.5-1.5B-Instruct, introduced in Section 7.1) without task-specific fine-tuning. The base model generated responses using the same prompt and inference configuration as the personalized system, but without any user-specific adaptation, which isolates the effect of user-specific LoRA personalization by fixing the underlying language model and generation settings. The difference between the personalized adapters and the non-personalized base model, measured across the four evaluation perspectives described in Section 6.7, provides an empirical measure of the effect of user-specific adaptation on

email-reply generation. The intent of this comparison is not to demonstrate superiority over all existing personalization approaches, but to examine whether user-specific LoRA adaptation improves measured stylistic alignment with a user's writing relative to the original base model, under identical experimental conditions and using automatic metrics alone; no human evaluation of overall reply quality was conducted in this study.

7.6 Data Leakage Prevention

The dataset size is modest relative to the scale of modern language models, so several measures were taken to reduce overfitting and memorization effects. Training, validation, and testing samples were separated at the email-reply pair level, validation loss was tracked throughout training, and the checkpoint with the lowest validation loss was selected for evaluation. Personalization was further constrained by using parameter-efficient LoRA adapters instead of full-model fine-tuning, which substantially reduced the number of tunable parameters. The reported results should accordingly be interpreted as an indication of personalization performance within a controlled experimental setup rather than as evidence of unconstrained generalization.

8 RESULTS AND DISCUSSION

8.1 Experimental Overview

EchoTwin was evaluated on a held-out test set of 223 email-reply pairs drawn from 10 selected users. The personalized LoRA adapters were compared against the underlying decoder-only Transformer large language model, which served as the non-personalized baseline; this comparison isolates the impact of user-specific adaptation while keeping the underlying language model and inference setup identical. Four complementary automatic evaluation metrics were used: lexical similarity (ROUGE), semantic similarity (Sentence-BERT), corpus-level style similarity, and stylometric analysis. Metrics were computed for each test sample and averaged over the complete held-out test set of 223 email-reply pairs; reported mean values represent overall average performance, and standard deviation values quantify the spread of metric scores across individual test samples. To assess the robustness of the observed improvements, a per-user analysis was performed by averaging metric scores for each user individually, and the Wilcoxon signed-rank test was applied to per-user ROUGE-L scores to determine whether the improvements obtained with the personalized LoRA adapters were statistically significant.

8.2 Lexical Similarity Evaluation

Lexical similarity was measured using ROUGE-1 F1 and ROUGE-L F1 [39]. These metrics capture lexical overlap between generated and reference responses: ROUGE-1 F1 measures unigram overlap, while ROUGE-L F1 reflects sentence-level similarity computed from the longest common subsequence between generated and reference replies. **Table 8** summarizes the evaluation results.

Table 8: ROUGE Evaluation and Statistical Summary

Metric	Base Mean	Base SD	Adapter Mean	Adapter SD	Abs. Improvement	% Improvement
ROUGE-1 F1	0.1123	0.088	0.1214	0.110	+0.0091	+8.10%
ROUGE-L F1	0.0731	0.046	0.0992	0.095	+0.0261	+35.65%

The personalized LoRA adapters achieved higher average scores than the base model on both lexical similarity metrics, with the larger gain on ROUGE-L (0.0731 to 0.0992, +35.65%) than on ROUGE-1 (+8.10%), indicating that user-specific adaptation improves both unigram overlap and the longer-range structural similarity captured by ROUGE-L. A Wilcoxon signed-rank test on the per-user average ROUGE-L scores confirmed that this improvement is statistically significant (statistic = 0.0, $p = 0.001953$) and unlikely to be due to chance [40].

The gain was also robust at the individual-user level: all 10 evaluated users showed higher average ROUGE-L after personalization, with no degradation for any user, and the median score rose alongside the mean (0.0721 to 0.0851), indicating that the improvement was distributed across the dataset rather than driven by a handful of high-performing samples. The personalized model's larger standard deviation (0.095 versus 0.046) is consistent with this picture rather than in tension with it: it reflects how widely emails vary in linguistic complexity and opportunity for personalization, not instability in the adaptation method. **Table 9** and **Figure 5** summarize these results.

Table 9: Statistical Analysis of ROUGE-L Performance

Analysis	Result
Number of test samples	223
Number of evaluated users	10
Baseline mean	0.0731
Personalized mean	0.0992
Baseline median	0.0721
Personalized median	0.0851
Wilcoxon signed-rank statistic	0.0
p-value	0.001953
Significance level	$\alpha = 0.05$
Users with improved performance	10/10 (100%)
Users with performance degradation	0/10 (0%)

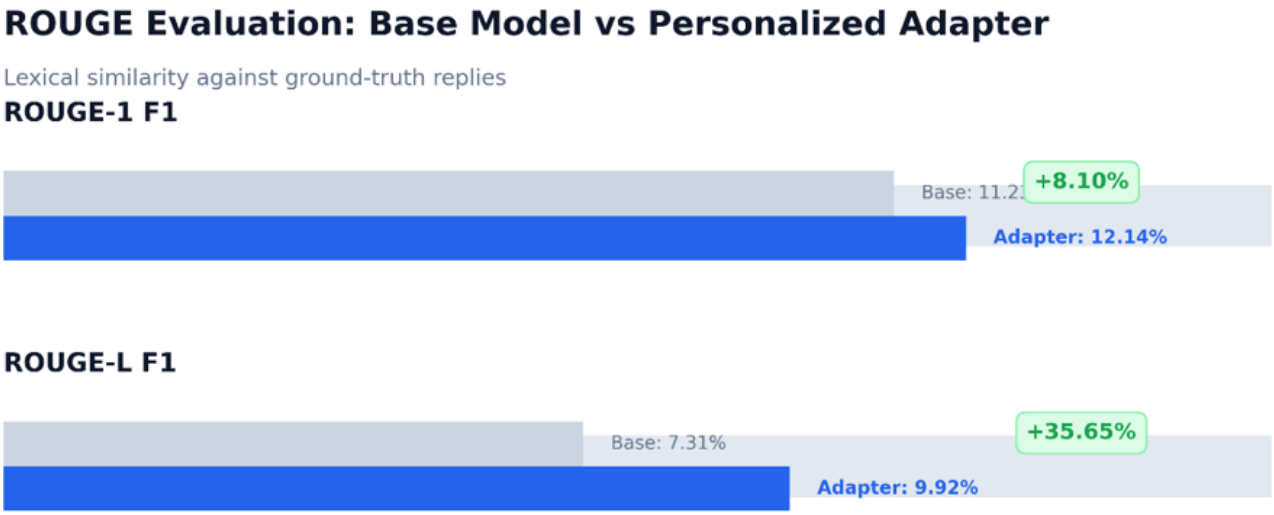

Figure 5: ROUGE-based lexical similarity comparison between the base model and personalized LoRA adapters

8.3 Semantic Similarity Evaluation

Sentence-BERT cosine similarity [32] was used to measure the semantic similarity between each generated reply and its corresponding reference reply based on sentence embeddings. Unlike lexical metrics, Sentence-BERT captures semantic relationships even when different wording is used to express similar meaning. **Table 10** presents the evaluation results.

Table 10: Semantic Similarity Evaluation Using Sentence-BERT

Metric	Base Mean	Base SD	Adapter Mean	Adapter SD	Abs. Difference	% Difference
Sentence-BERT Similarity	0.2626	0.1829	0.2356	0.1713	−0.0270	−10.29%

The personalized LoRA adapters achieved a lower average Sentence-BERT similarity to the reference reply than the non-personalized baseline (0.2626 to 0.2356, a decrease of 10.29%), with comparable standard deviations across the two models (0.1829 versus 0.1713). A Wilcoxon signed-rank test on the paired scores across all 223 evaluation samples confirmed that this difference is statistically significant (statistic = 10346.0, $p = 0.02637$) [40].

This decrease should not be read as inferior generation quality: Sentence-BERT measures similarity to a single reference reply, whereas the goal of personalization is to reflect each user's own historical style, and lexical and structural changes that improve stylistic fit can reduce embedding-based similarity to that one reference even when the response remains coherent. The per-user breakdown supports this reading, with seven of ten users showing lower semantic similarity after personalization and three showing improvement, indicating that the effect varies across users rather than acting uniformly. **Figures 6–7** and **Table 11** report the full statistical summary and per-user and aggregate comparisons; this trade-off is

examined further below through corpus-level style similarity and stylometric analysis, which assess stylistic alignment directly rather than similarity to a single reference.

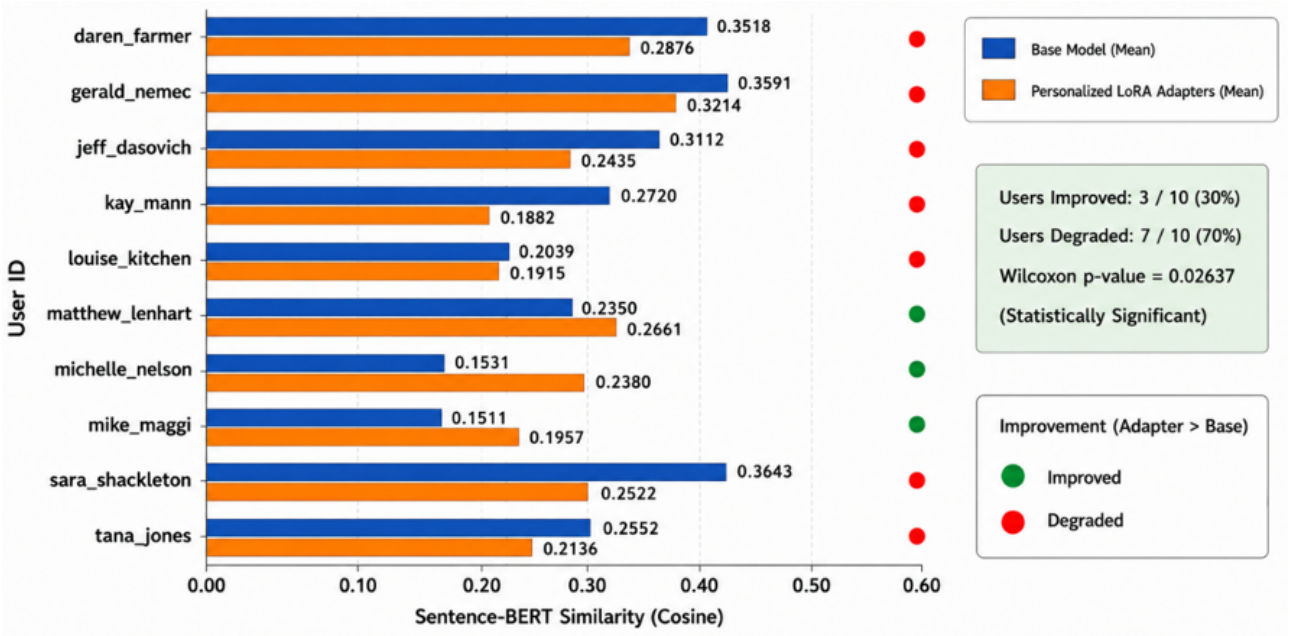

Figure 6: Per-user Sentence-BERT semantic similarity comparison between the base model and personalized LoRA adapters

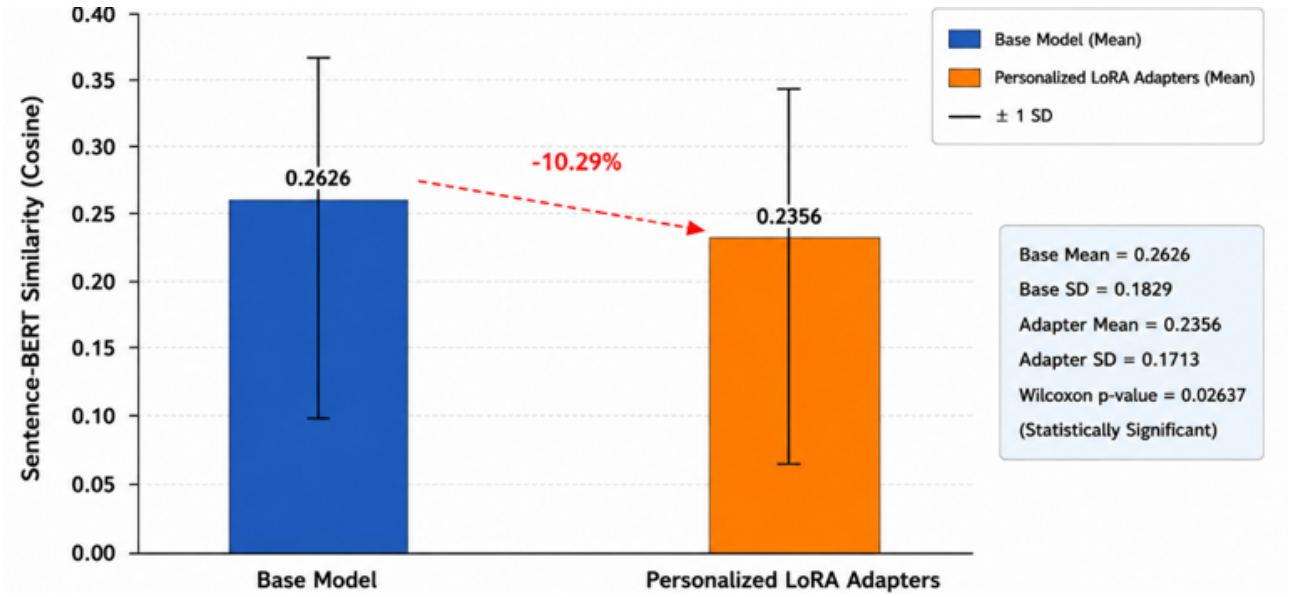

Figure 7: Sentence-BERT semantic similarity comparison between the base model and personalized LoRA adapters

8.4 User Corpus Style Similarity

Corpus-level style similarity was computed by comparing generated replies with the historical reply corpus of the corresponding target user using sentence embedding similarity [32]. Unlike the semantic similarity evaluation in Section 8.3, which compares each generated reply to its individual reference reply, this evaluation measures the degree of similarity between generated responses and the general writing style of the user's historical email corpus. **Table 12** summarizes the results.

These results show that personalization increased the degree of matching with user-specific communication patterns observed in the historical email corpus. The personalized adapters achieved an average style similarity score of 0.3730, compared with 0.3531 for the base model, a 5.64% improvement, suggesting that the personalized LoRA adapters adapted generated responses to the communication patterns present in the historical user corpus, resulting in stronger stylistic

Table 11: Statistical Analysis of Sentence-BERT Similarity

Analysis	Result
Number of test samples	223
Number of evaluated users	10
Baseline mean	0.2626
Personalized mean	0.2356
Baseline SD	0.1829
Personalized SD	0.1713
Wilcoxon signed-rank statistic	10346.0
p-value	0.02637
Significance level	$\alpha = 0.05$
Users with improved performance	3/10 (30%)
Users with lower performance	7/10 (70%)

Table 12: User Corpus Style Similarity Evaluation

Metric	Base Mean	Base SD	Adapter Mean	Adapter SD	Abs. Improvement	% Improvement
User Corpus Style Similarity	0.3531	0.190	0.3730	0.164	+0.0199	+5.64%

agreement with past user responses. This is further supported by the standard deviation: the personalized adapters exhibited a lower standard deviation (0.164) than the base model (0.190), implying more consistent stylistic behavior across email samples and users. These findings suggest that personalization increased both the average level of stylistic alignment and the consistency of reproducing user-specific communication patterns in generated responses.

Figure 8 presents user-level style similarity results for the base model and the personalized LoRA adapters. Most users showed improvement after personalization, indicating that the adapters effectively adapted to user-specific communication patterns observed in the training data. The largest improvement was observed for mike_maggi (+0.1158), followed by daren_farmer (+0.0610) and michelle_nelson (+0.0583). A small number of users experienced decreased performance, but the overall trend indicates that user-specific LoRA adapters are more likely than not to improve stylistic alignment with past user replies. The variation observed across users further suggests that the effectiveness of personalization depends on the number, consistency, and diversity of available training examples.

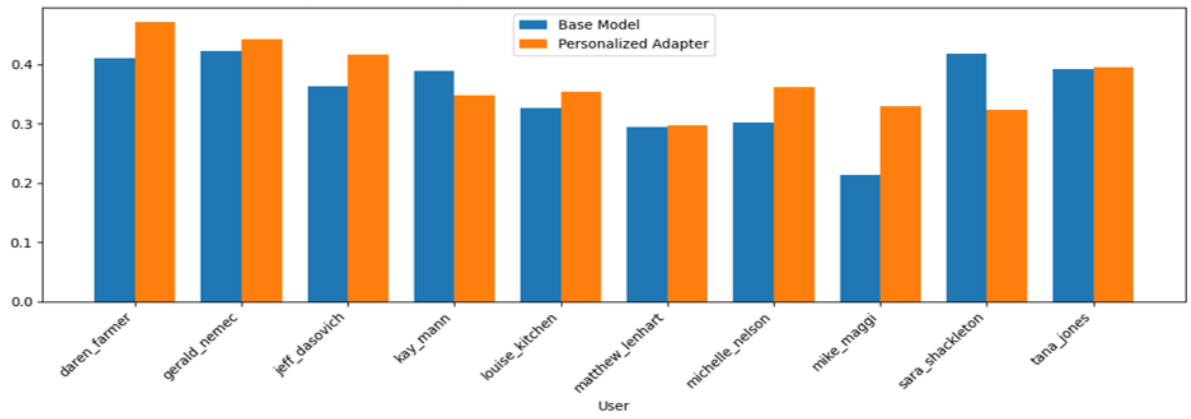

Figure 8: Per-user style similarity comparison between the base model and personalized LoRA adapters

8.5 Stylometric Analysis

Stylometric similarity was evaluated using the mean absolute stylometric delta between each generated reply and the corresponding historical writing style of the target user. This delta summarizes differences across multiple writing characteristics, including reply length, sentence structure, punctuation usage, and greeting and closing expressions, with lower delta values indicating closer alignment with the user's historical writing style. **Table 13** summarizes the evaluation

results.

Table 13: Stylometric Delta Evaluation

Metric	Base Mean	Base SD	Adapter Mean	Adapter SD	Delta Reduction	% Reduction
Mean Absolute Stylometric Delta	0.9315	0.4744	0.6962	0.4218	0.2353	25.26%

These results demonstrate that the personalized LoRA adapters substantially reduce stylometric distance between generated responses and each user's historical writing style. The average stylometric delta decreased from 0.9315 to 0.6962, a 25.26% reduction, indicating that personalizing generated style produces replies that are stylistically closer to a user's past communications, since lower stylometric delta values represent higher stylistic similarity.

A Wilcoxon signed-rank test on the stylometric delta values, applied at the case level across 223 paired samples, produced a statistic of 20715.0 with a p-value of 7.36×10^{-18} [40], indicating an extremely significant improvement. A second Wilcoxon signed-rank test at the user level, across 10 paired users, produced a statistic of 55.0 with a p-value of 0.000977, confirming that the observed improvement remained statistically significant across users.

A per-user analysis further supports the robustness of the proposed personalization strategy. All 10 evaluated users achieved lower average stylometric delta values after user-specific LoRA adaptation, a 100% improvement rate with no degradation observed for any user. The median stylometric delta decreased from 0.8543 for the baseline model to 0.5935 for the personalized model; the simultaneous decrease in mean and median values indicates that the observed improvement was consistent across the evaluation dataset rather than driven by a small number of favorable cases.

The personalized model also exhibited a lower standard deviation (0.4218) than the baseline model (0.4744), suggesting that personalization improved not only stylistic alignment but also the consistency of stylometric behavior across the evaluation dataset. An outlier analysis using the interquartile range (IQR) identified only 6 outlier cases (2.69%) among the 223 analyzed samples; excluding these outliers, the personalized adapters still reduced the stylometric delta by 27.72%, and the Wilcoxon signed-rank test remained statistically significant ($p < 0.001$), indicating that the observed improvement is not driven by a small number of extreme observations.

Table 14 provides additional statistical evidence for the effectiveness of the proposed personalization approach. Significant improvements were observed at both the case and user levels, with every evaluated user benefiting from user-specific LoRA adaptation. The reduction in both mean and median stylometric delta values, together with the robustness of the result after outlier removal, confirms that the observed improvement is consistent and is not driven by a small number of extreme samples.

Table 14: Statistical Summary of Stylometric Evaluation

Analysis	Result
Number of test samples	223
Number of evaluated users	10
Baseline mean delta	0.9315
Personalized mean delta	0.6962
Baseline median delta	0.8543
Personalized median delta	0.5935
Baseline SD	0.4744
Personalized SD	0.4218
Delta reduction	0.2353
Percentage reduction	25.26%
Improved users	10/10 (100%)
Users with performance degradation	0/10 (0%)
Case-level Wilcoxon statistic	20715.0
Case-level p-value	7.36×10^{-18}
User-level Wilcoxon statistic	55.0
User-level p-value	0.000977
Outlier cases	6/223 (2.69%)
Robustness after outlier removal	Improvement remained significant ($p < 0.001$)

Figure 9 illustrates the reduction in mean stylometric distance achieved by the personalized LoRA adapters relative to the

baseline model, indicating improved alignment with user-specific writing characteristics; a lower delta value indicates a closer match to the target user's historical writing style.

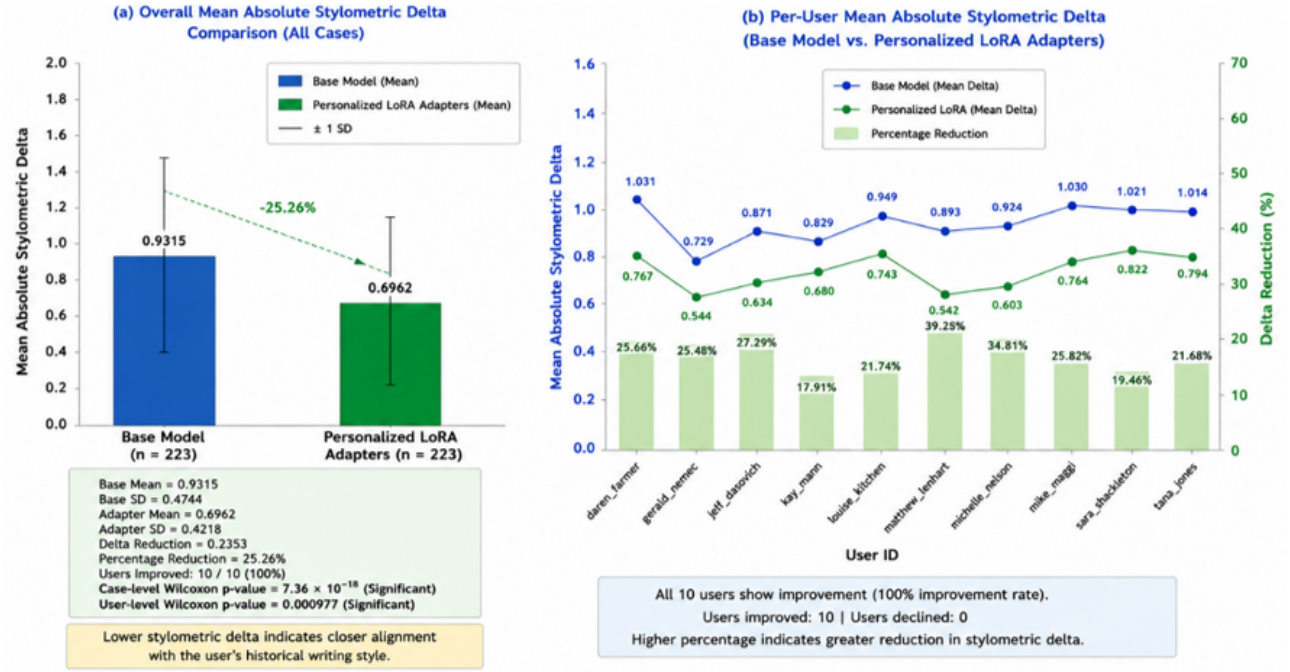

Figure 9: Stylometric delta reduction achieved by personalized LoRA adapters

8.6 Overall Performance Summary

Experimental results indicate that user-specific LoRA adaptation can effectively improve measured stylistic alignment in personalized email-reply generation, though not uniformly across every evaluation perspective. Lexical evaluation showed a statistically significant improvement over the non-personalized baseline, while semantic evaluation revealed a statistically significant decrease in Sentence-BERT similarity, consistent with the expected trade-off between maximizing embedding-based similarity to a single reference reply and adapting responses to an individual's writing style. Style similarity and stylometric analysis at the corpus level further demonstrate the ability of the proposed method to capture user-specific writing characteristics. Together with the statistical significance tests and per-user analyses reported in the preceding subsections, these results suggest that the observed changes are consistent across the evaluated users and are unlikely to be explained by random variation. These findings should be read as evidence of improved stylistic alignment as measured by automatic metrics; they do not, by themselves, establish an improvement in overall reply quality or usefulness, which would require human evaluation.

Table 15 summarizes the main experimental findings across all four evaluation perspectives.

9 LIMITATIONS AND FUTURE WORK

9.1 Gmail-Only Integration

The current implementation uses Google OAuth 2.0 and the Gmail API to integrate with Gmail for secure email retrieval, draft generation, and response submission. The proposed framework does not currently integrate with other email providers, such as Microsoft Outlook, Yahoo Mail, or enterprise email platforms, and is therefore applicable only to Gmail users at present. Future work will extend compatibility to additional providers through provider-specific APIs and standardized email protocols.

9.2 Cold-Start Personalization

The proposed personalization method learns writing style from a user's past email responses, so the framework may not fully capture an individual's communication style when only a small number of user-specific examples are available, a limitation commonly referred to as the cold-start problem in personalized systems. Future work may address this challenge through

Table 15: Summary of Experimental Findings

Evaluation Aspect	Main Finding
Lexical similarity (ROUGE)	User-specific LoRA adaptation improved lexical similarity relative to the non-personalized baseline.
Semantic similarity (Sentence-BERT)	A modest decrease was observed, reflecting the trade-off between semantic similarity and stylistic personalization.
Corpus-level style similarity	Personalized replies more closely matched the historical writing style of the target users.
Stylometric analysis	Personalized replies showed stronger alignment with users' writing characteristics, with statistically significant improvements across the evaluation dataset.
Overall assessment	The proposed approach consistently improved measured stylistic alignment with individual users, with a corresponding trade-off in reference-based semantic similarity; automatic metrics alone do not establish an improvement in overall reply quality, which was not assessed by human raters in this study.

retrieval-augmented personalization, explicit modeling of user preferences, approved-reply histories, and mechanisms for adaptively refining user profiles. Although the users selected for this study provided sufficient training examples to run the reported experiments, the available data remained relatively small compared with the scale typically used to train modern large language models. The findings reported in this study should therefore be interpreted in the context of the dataset and experimental setting used for evaluation, rather than as evidence of universal generalizability.

9.3 Dataset and Generalization Limitations

Evaluation was performed on a subset of 2,221 email-reply pairs from the Enron Email-Reply corpus, collected from 10 selected users. While these users provided sufficient data to conduct user-specific personalization experiments, the dataset captures only a narrow range of communication styles and application domains, so the results should be interpreted in the context of the evaluated dataset and may not directly generalize to broader real-world email environments. Larger and more diverse datasets, a wider range of users, and cross-domain evaluation represent promising directions for further assessing the generalizability of the proposed approach.

9.4 Model Adaptation Limitations

The proposed framework generates context-aware personalized email replies, but response quality still depends on the information available in the email input and in the user-specific training data. Generated responses may therefore be incomplete or contain erroneous inferences or stylistic inconsistencies, particularly when input emails are ambiguous, lack sufficient context, or require domain-specific knowledge that is absent from the training data. In the current implementation, individual email messages are processed with limited conversational context; the system does not explicitly model advanced conversation-thread reasoning, attachment understanding, external knowledge retrieval, or long-term user memory, and such constraints can affect response quality in complex communicative situations. Future work could address these challenges through improved email intent understanding, richer modeling of conversation history, retrieval-augmented generation, attachment-aware reasoning, and more sophisticated contextual personalization techniques.

9.5 Evaluation Limitations

Evaluation was conducted on held-out samples from the Enron Email-Reply dataset in a controlled offline setting, using several complementary automatic metrics, including lexical similarity, semantic similarity, corpus-level style similarity, and stylometric analysis. These metrics do not fully capture aspects of personalized communication such as long-term user satisfaction, editing effort, productivity gains, trust, or real-world adoption, and the experiments were performed with a limited number of users on a single public email dataset, which may constrain the generalizability of the reported results. Future work should investigate larger and more diverse datasets, additional evaluation benchmarks, cross-domain validation, and longitudinal user studies to further assess the robustness and generalizability of the proposed personalization approach.

9.6 Future Research Directions

The proposed personalization framework can be extended in several directions. First, privacy-preserving personalization techniques, including federated learning, secure aggregation, and differential privacy, have not been applied or tested in the current study and represent promising directions for future research [20, 21, 22]. Second, personalization quality could be further improved through retrieval-augmented generation, improved style modeling, and long-term user feedback mechanisms. Third, the proposed framework could be made more broadly applicable by supporting additional languages, email providers, and communication platforms. Finally, future work could explore applying the proposed personalization approach to customer support systems, business messaging, and collaborative communication platforms, while preserving privacy guarantees and human-approval mechanisms.

10 CONCLUSION

This paper introduced EchoTwin, a personalized email-reply generation framework that integrates Gmail access, user-specific LoRA adaptation, and style-aware response generation while preserving user control over generated replies. The proposed framework addresses the question of how effectively parameter-efficient personalization can generate email responses that better reflect individual writing styles. The framework was evaluated on a curated subset of the Enron Email-Reply dataset comprising 2,221 email-reply pairs from 10 selected users, with experimental evaluation conducted on 223 held-out test samples comparing user-specific LoRA adapters with a non-personalized Transformer large language model baseline. The results show consistent improvements in lexical similarity, corpus-level style similarity, and stylometric alignment, with the proposed method achieving improvements of 8.10% in ROUGE-1, 35.65% in ROUGE-L, and 5.64% in corpus-level style similarity, together with a 25.26% reduction in mean stylometric distance. Sentence-BERT similarity decreased slightly, a consequence of the expected trade-off between maximizing semantic similarity and adapting responses to individual writing styles. Statistical significance tests and per-user analyses further confirm that the observed improvements are consistent across the evaluated users and are unlikely to be explained by random variation. In summary, the experimental results indicate that user-specific LoRA adaptation is an effective approach for improving measured stylistic alignment in email-reply generation, at the cost of a modest, statistically significant reduction in reference-based semantic similarity. The study was conducted on a relatively small dataset with a limited number of users and relied exclusively on automatic evaluation metrics, so the reported findings should be interpreted within the evaluated experimental setting, and as evidence of stylistic alignment rather than of broader gains in reply quality or of generalizability beyond the evaluated users. Future work will explore larger and more diverse datasets, wider user populations, more sophisticated personalization strategies, privacy-preserving adaptation methods, and support for additional email providers and communication platforms.

AUTHOR CONTRIBUTION STATEMENT

All authors contributed equally to the study conception and design. Material preparation, data collection, and analysis were performed by the authors. The first draft of the manuscript was written by the authors, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This study did not involve human participants or animals. Ethical approval and consent to participate are therefore not applicable.

CONSENT FOR PUBLICATION

Not applicable.

DATA AVAILABILITY

The data underlying this study are drawn from the Enron Email-Reply dataset, a curated subset of the original Enron email corpus, both of which are publicly available: the Enron Email-Reply dataset is distributed on Kaggle [29], and the original Enron corpus is distributed by Carnegie Mellon University [30].

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to Dr. Mohammed Salah for his valuable supervision, guidance, and continuous support throughout the development of the EchoTwin project. His insightful comments and guidance contributed significantly to improving the system design and overall structure of this work. The authors also gratefully acknowledge Anasimon Samir, Jasmin Eshak, and Merola Akram for their valuable assistance, collaboration, and constructive contributions during the project's development and evaluation phases. Their support was instrumental in implementing the system and refining its schematics. The authors further thank the Faculty of Computer Science at El-Nahda University in Beni-Suef for providing the environment and support. The authors also acknowledge the use of ChatGPT-5 to assist in refining the manuscript. All AI-generated content was reviewed and edited by the authors, who take full responsibility for the final version.

FUNDING

This research received no external funding.

DISCLOSURE STATEMENT

The authors declare that they have no competing interests.

REFERENCES

[1] McKinsey Global Institute, "The social economy: Unlocking value and productivity through social technologies," tech. rep., McKinsey & Company, New York, NY, USA, 2012.

[2] M. Plummer, "How to spend way less time on email every day," *Harvard Business Review*, vol. 22, 2019.

[3] A. Kannan, K. Kurach, S. Ravi, T. Kaufmann, A. Tomkins, B. Miklos, G. Corrado, L. Lukacs, M. Ganea, P. Young, *et al.*, "Smart reply: Automated response suggestion for email," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 955–964, 2016.

[4] M. X. Chen, B. N. Lee, G. Bansal, Y. Cao, S. Zhang, J. Lu, J. Tsay, Y. Wang, A. M. Dai, Z. Chen, *et al.*, "Gmail smart compose: Real-time assisted writing," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2287–2295, 2019.

[5] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[7] Google Developers, "Using OAuth 2.0 to access Google APIs." `https://developers.google.com/identity/protocols/oauth2`, 2026. Accessed: May 11, 2026.

[8] Google Developers, "Gmail API Reference: `users.messages.send`." Google for Developers. [Online]. Available: `https://developers.google.com/gmail/api/reference/rest`, 2026. Accessed: May 11, 2026.

[9] OWASP Foundation, "OWASP Top Ten Web Application Security Risks." OWASP. [Online]. Available: `https://owasp.org/www-project-top-ten/`, 2025. Accessed: May 12, 2026.

[10] E. Tabassi, "Artificial intelligence risk management framework (ai rmf 1.0)," Tech. Rep. NIST AI 100-1, National Institute of Standards and Technology, Gaithersburg, MD, USA, 2023.

[11] A. Silva, P. Tambwekar, and M. Gombolay, "Fedperc: Federated learning for language generation with personal and context preference embeddings," in *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 869–882, 2023.

[12] A. Salemi, S. Kallumadi, and H. Zamani, "Optimization methods for personalizing large language models through retrieval augmentation," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 752–762, 2024.

[13] A. Salemi and H. Zamani, "Comparing retrieval-augmentation and parameter-efficient fine-tuning for privacy-preserving personalization of large language models," in *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, pp. 286–296, 2025.

[14] Z. Tan, Q. Zeng, Y. Tian, Z. Liu, B. Yin, and M. Jiang, "Democratizing large language models via personalized parameter-efficient fine-tuning," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6476–6491, 2024.

[15] C. Sun, K. Yang, R. G. Reddy, Y. Fung, H. P. Chan, K. Small, C. Zhai, and H. Ji, "Persona-db: Efficient large language model personalization for response prediction with collaborative data refinement," in *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 281–296, 2025.

[16] Y. Xu, J. Zhang, A. Salemi, X. Hu, W. Wang, F. Feng, H. Zamani, X. He, and T.-S. Chua, "Personalized generation in large model era: A survey," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24607–24649, 2025.

[17] Y. Li, Z. Tan, P. Branco, and Y. Liu, "Privacy-preserving parameter-efficient fine-tuning for large language model services," *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[18] Y. Wang, Y. Lin, X. Zeng, and G. Zhang, "Privatelora for efficient privacy preserving llm," *arXiv preprint arXiv:2311.14030*, 2023.

[19] Z. Zeng, J. Wang, J. Yang, Z. Lu, H. Li, H. Zhuang, and C. Chen, "Privacyrestore: Privacy-preserving inference in large language models via privacy removal and restoration," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10821–10855, 2025.

[20] R. Singhal, K. Ponkshe, and P. Vepakomma, "Fedex-lora: Exact aggregation for federated and efficient fine-tuning of large language models," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1316–1336, 2025.

[21] Z. Shen, J. Lu, H. Wan, and J. Chen, "Sdflora: Selective decoupled federated lora for privacy-preserving fine-tuning with heterogeneous clients," *arXiv preprint arXiv:2601.11219*, 2026.

[22] J. Liu, Y. Miao, N. Xi, and J. Liu, "Rethinking lora for privacy-preserving federated learning in large models," *arXiv preprint arXiv:2602.19926*, 2026.

[23] A. Salemi, S. Mysore, M. Bendersky, and H. Zamani, "Lamp: When large language models meet personalization," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7370–7392, 2024.

[24] I. Kumar, S. Viswanathan, S. Yerra, A. Salemi, R. A. Rossi, F. Dernoncourt, H. Deilamsalehy, X. Chen, R. Zhang, S. Agarwal, *et al.*, "Longlamp: A benchmark for personalized long-form text generation," *arXiv preprint arXiv:2407.11016*, 2024.

[25] B. Jiang, Z. Hao, Y.-M. Cho, B. Li, Y. Yuan, S. Chen, L. Ungar, C. J. Taylor, and D. Roth, "Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale," *arXiv preprint arXiv:2504.14225*, 2025.

[26] X. Fu, H. A. Rahmani, B. Wu, J. Ramos, E. Yilmaz, and A. Lipani, "Pref: Reference-free evaluation of personalised text generation in llms," *arXiv preprint arXiv:2508.10028*, 2025.

[27] J. Park, D. Kim, and T. Moon, "Prisp: Privacy-safe few-shot personalization via lightweight adaptation," in *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24986–25003, 2026.

[28] X. Li, R. Zhou, Z. C. Lipton, and L. Leqi, "Personalized language modeling from personalized human feedback," *arXiv preprint arXiv:2402.05133*, 2024.

[29] Kaggle, "Enron email-reply dataset." Kaggle. [Online]. Available: `https://www.kaggle.com/datasets/oanannv/enron-email-reply-dataset`, 2023. Accessed: May 17, 2026.

[30] Carnegie Mellon University, "Enron email dataset." CMU School of Computer Science. [Online]. Available: `https://www.cs.cmu.edu/~enron/`, 2015. Accessed: May 17, 2026.

[31] N. Sakimura, J. Bradley, and N. Agarwal, "Proof key for code exchange by oauth public clients," Tech. Rep. RFC 7636, RFC Editor, Sept. 2015.

[32] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

[33] P. Juola, "Authorship attribution," *Foundations and Trends® in Information Retrieval*, vol. 1, no. 3, pp. 233–334, 2008.

[34] M. Koppel, J. Schler, and S. Argamon, "Authorship attribution in the wild," *Language Resources and Evaluation*, vol. 45, no. 1, pp. 83–94, 2011.

[35] React Documentation, "React documentation." Online, 2026. Available: `https://react.dev/learn`. Accessed: May 17, 2026.

[36] S. Tiangolo, "Fastapi documentation," *URL: https://fastapi. tiangolo. com*, 2025.

[37] B. Ward, *SQL Server 2025 Unveiled: The AI-Ready Enterprise Database with Microsoft Fabric Integration*. Berkeley, CA, USA: Apress, 2025.

[38] H. Face, "Transformers documentation," *URL: https://huggingface. co/docs/transformers/index*, 2024.

[39] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, pp. 74–81, 2004.

[40] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.